

# SCIENTOSEMANTICS OF APPLIED RESEARCH USING AI AND MACHINE LEARNING – A SEQUENTIAL APPROACH

Gustaf Nelhans and Johan Eklund,

Data as Impact Lab

2021-04-19



THE SWEDISH SCHOOL OF LIBRARY  
AND INFORMATION SCIENCE  
UNIVERSITY OF BORÅS

## Contents

Introduction - objectives and background.....	3
Scientometric analyses .....	4
A note about “exploratory scientometrics” .....	5
Methodology.....	5
Scientometric report, Air Quality Research .....	7
Results .....	7
Document type and publication year .....	7
Authorship.....	9
Organisation level .....	11
Journal level .....	14
Co-citation analysis.....	16
Keywords and co-word analysis .....	17
Scientometric report, Traffic safety research .....	20
Results .....	20
Document type and publication year .....	20
Authorship.....	22
Organisation level .....	23
Journal level .....	26
Co-citation analysis.....	27
Keywords and co-word analysis .....	29
Discussion.....	32
A scientosemantic approach to bibliographic data.....	33
Question answering in abstracts .....	34
Search engine for ranked retrieval of abstracts.....	35
Conclusions .....	38
References .....	39

## Introduction - objectives and background

This report presents the results of a combined scientometric and machine learning exercise with the objective to help in the description of relevant research within the two topics of Air Quality research and Traffic safety research in relation to the application of artificial intelligence and machine learning techniques. It was performed by the Data as Impact Lab at the Swedish School of Library and information science, University of Borås.

The assignment consisted of investigating research where AI and advanced data analysis are used to study the respective subject areas of air quality issues and traffic safety. Specifically, we were asked to identify the development of research in the respective areas during recent years. Contexts that were requested related to which universities/areas/countries are active and identifying helpful review articles. There was also an interest in determining which data sources and methodological approaches have been used in the studies. For example, we were supplied with issues such as analysing air quality, distribution of emissions, and exposure to emissions in humans. The data sources could vary significantly from measuring emissions levels in measuring stations to satellite data and mobile phone data to measuring movement patterns.

The first stage of the work related to identifying meaningful terms to use in the searches for relevant research in the publication databases used. Clarivate Web of Science was chosen for the task, one of the most comprehensive publication databases with extensive quality control. It also contains bibliographical references to other scholarly research, which is a prerequisite for doing citation analysis.

Most publication databases employ a classic search interface. It means that data is first retrieved and is then stored for analysis in a sequential manner.<sup>1</sup> We first search for relevant literature using specific terms covering the subject in question and then employ various scientometric methods to investigate the content and metadata of the literature that we find. The client supplied these search terms. They consisted of subject terms such as “air quality”, “atmospheric pollution”, and “particle emission” for the study of Air Quality research. For Traffic Safety research, the terms included “traffic safety”, “road safety”, “safety assessment, and “traffic accidents”. These terms were combined with methodological terms used to describe current AI approaches within research such as “artificial intelligence”, “machine learning”, “deep learning”, and more specified terms as “random forest” and “support vector machine”. In order to transform these terms into Boolean search strings, several approaches need to be performed. It involves combining words into concepts, joining strings of terms within quotes (e.g. “air quality”), combining different concepts using the OR operator and combining them to identify the Union of two sets using the AND operator.

---

<sup>1</sup> Another approach, using live API access to publication data using the openly available CrossRef database as a source and Open Citations for linking references between the literature was investigated but rejected based on time constraints in the assignment and coverage. Such an exploratory approach would have yielded more emphasis on the acquisition data instead of a focus on analysis. Another issue is that not all publishers deliver citation data to CrossRef or do not allow its use, which means that novelty of data would be substituted for quality control. Though, it would be interesting to develop a live interface to live metadata for research, especially in a field such as AI/machine learning, where development is very quick.

In some cases, instead of using the less strict AND operator to combine terms that need to be close to each other, the NEAR/4 operator was used, meaning that if two terms are found within a four-word radius, they will be retrieved together (e.g. “*air* and water *quality*”). While it is important to yield a large enough data set to work with, no search string is perfect. Therefore, after an initial trial and error period, the effort most often ends with a reasonably wide string to identify as many relevant documents as possible while, at the same time, block off irrelevant terms based on the concept of precision and recall. At some point, introducing new terms does not yield a significant change in the number of hits. This coincides with the search string being “saturated”. The individual search strings that were found meaningful for each topic are found at the beginning of each part of the study.

## Scientometric analyses

An overview of each subject matter is presented in the following two chapters. The study uses a traditional bibliometric methodology to identify and aggregate bibliographic information from the Web of Science. Apart from Title, abstract and keywords found in each article that is retrieved, we are also able to extract metadata about authors, their affiliations and even funding data for the research (though, the latter is not used in the analysis. Lastly, citation data based on the bibliographic references of each article is retrieved from Web of Science).

Together, these different sources of information can be aggregated using scientometric methodology. Based on the concept of citation analysis, aggregation of data at different entity levels: article, author, source (e.g. the journal wherein the article is published in), or based on organisation data (university, department or country), three different scientometric methods are employed:

1. Co-authorship, meaning that authors or organisations are linked together if their respective entities are linked together based on authorship.
2. Bibliographic coupling, wherein two entities are linked together if they cite the same references as sources for the research, and
3. Co-citation analysis, where two cited sources (documents, authors or source journals) are linked together if the same entity cites them.

Additionally, a text mining approach, called co-word analysis, links relevant noun phrases to each other if found in titles and abstracts in the data set. The text-based analysis of keywords and key terms in the WoS dataset’s titles and abstracts is used. Keywords are registered at the article level by the publisher, often chosen by researchers themselves but sometimes chosen from a list of pre-determined keywords. This algorithm considers pair-wise relationships between all keywords identified in the articles citing the institutes’ publications based on how often the terms occur together in the “author generated” keyword list.<sup>2</sup> Another way of identifying key terms and phrases uses terms identified in the articles’ titles and abstracts. This is a more “free form” of text, and while sometimes noisy, can provide insights in the actual terminology used instead of the more restricted set of keywords. Using VOSviewer, the co-word algorithm filters the text for meaningful noun phrases, including nouns and

---

<sup>2</sup> As opposed to Keywords PLUS™, which is a set of keywords that is added by Web of Science.



adjectives in front of nouns to identify semantic phrases of relevance, using linguistic techniques.

## A note about “exploratory scientometrics”

As opposed to well-known scientometric uses for evaluative purposes, we are not interested in evaluating or ranking research per se but instead exploring the data generated and finding interesting patterns and aspects of the data to investigate further. In this study, most presented data is based on the notion of “exploratory scientometrics”. This means that, instead of focusing on the ranking of entities for analysis, we try to convey the relational aspects of the scholarly papers found in the original searches. Be it citations or similar use of terminology, instead of top-10 lists; we try to show who collaborates with whom, the overlap between research interests at one organisation with another, as well as the similarities in terminology found between different levels of analysis. Therefore, the preferred means of exploring the results are from network visualisations of the scientometric data, which has the advantage of conveying much information in a condensed format. It also allows the user to explore the results themselves. Therefore, we only give some hints about interpreting the results and leaving it to the expert reader to convey meaning and conclusions about what is found.

We intend that the following analyses and illustrations of research publications in the field should give options for identifying the research’s breadth and depth, as seen through the lens of scientometrics. Moreover, it affords “hypothesis generation” options to explore the vast set of data and find new insights into the work. Therefore, we intend to provide the reader with maps of the landscape and hints at interpreting the results. However, we intend that most of the actual analysis and conceptualisation will be done by the reader.

A final word about the visualisations shown in the report. Since the flat format of a report is somewhat limiting for a detailed analysis of the data, we also provide all visualisations in the report in an online appendix where the graphics are shown in a larger size and pdf format for vectorised versions that could be zoomed into.

## Methodology

For the downloaded set of publications from WoS, the following report shows the most relevant data. All data is based on full counts at the document level. No fractionalisation or field normalisation is performed in tabular data. We perform contributor fractionalisation in the visualisations when relevant.

Tabular data is based on Web of Science data. Preparation of data was made using HistCite<sup>3</sup>, a legacy software developed by Dr Eugene Garfield (1925-2017). To illustrate the bibliographic data and aggregate it so that more comprehensive information can be elucidated, tabular data is often accompanied by bibliographic visualisations. A software package, VOSviewer<sup>4</sup> (van Eck & Waltman, 2010), was used for most visualisations. It is created by

---

<sup>3</sup> A legacy version is available for Windows computers:

[https://support.clarivate.com/ScientificandAcademicResearch/s/article/HistCite-No-longer-in-active-development-or-officially-supported?language=en\\_US](https://support.clarivate.com/ScientificandAcademicResearch/s/article/HistCite-No-longer-in-active-development-or-officially-supported?language=en_US)

<sup>4</sup> <https://www.vosviewer.com/>

researchers at the Centre for Science and Technology Studies at Leiden University. As opposed to generic visualisation software, it has been designed to read output data files from citation databases such as Clarivate Web of Science and Elsevier Scopus, alleviating the often burdensome handling of these nested data frames. Additional handling of data was performed using R, Python and MS Excel.

# Scientometric report, Air Quality Research

(All data were downloaded from Web of Science Core Collection on 2021-01-18)

**Timespan:** open time window. **Indexes:** SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI.

## Search string:

*TS= ((air OR atmospher\* OR aerosol\* OR particl\*) NEAR/4 (qual\* OR pollut\* OR emiss\* OR expos\*))*

*AND*

*TS= (("machine learning" OR "deep learning" OR "artific\* intel\*" OR "neural network\*" OR "support vector machine\*" OR "reinforcement learning" OR "random forest\*"))*

**Identified documents:** 3.166

## Results

### Document type and publication year

For the publications identified in Web of Science, we show the number of documents published on a yearly basis (Figure 1). We find the number of articles per year is increasing, especially after 2014. Since no strict criteria for limiting the inclusion were performed, the citation database included documents published in 2021 and a few with an “unknown” date. These are generally preprints without a version of record that have yet to receive a publishing date.

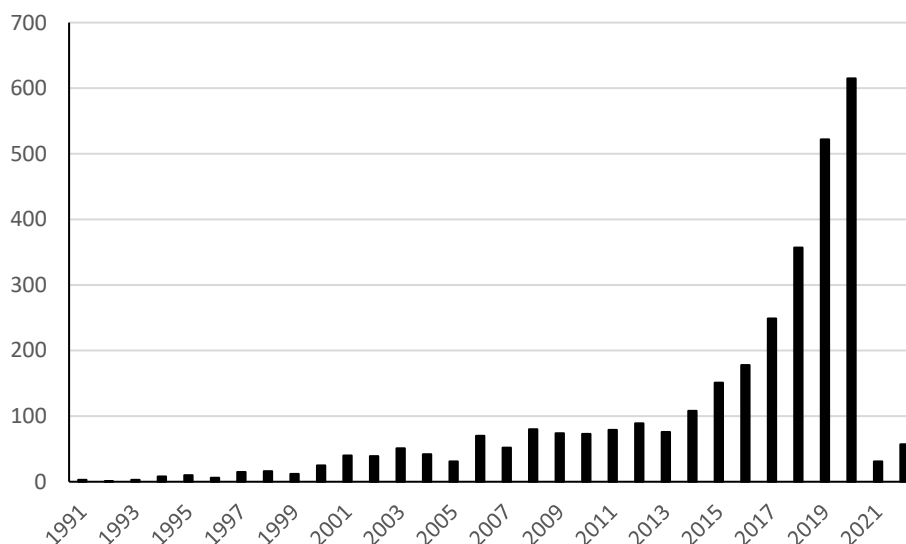


Figure 1. The number of yearly publications since 1991. The last column shows “unknown year, which primarily relates to ahead-of-print publications.

Most documents are of the category (peer reviewed) “article” (Table 1). Quite a large share of the documents are proceedings papers, which is quite common in engineering sciences. There is also a small number, but in terms of citations, quite significant publications of the review kind. Other document types were relatively few. Although some of these are not peer reviewed, it was deemed unnecessary to remove these publications since they might still be relevant for information purposes. As with self-citations, as will be seen later, when using bibliographic data for exploratory tasks, less strict inclusion criteria is often warranted. Apart from publication numbers, two additional columns are shown. The last one, GCS, stands for Global Citation Score and corresponds to the number of citations the documents at a particular year has received up to the time of retrieving data for the study. The middle one, LCS, stand for “Local Citation Score” and shows the number of citations the publications of a particular year has received *within* the data set.

Table 1: Document type

DOCUMENT TYPE	RECS	PERCENT	TLCS	TGCS
Article	2169	68.6	8907	34101
Proceedings Paper	750	23.7	478	2016
Review	92	2.9	188	2316
Article; Proceedings Paper	81	2.6	387	2125
Article; Early Access	55	1.7	0	14
Software Review	6	0.2	30	43
Article; Data Paper	4	0.1	0	15
Meeting Abstract	3	0.1	0	0
Review; Early Access	2	0.1	0	1
Editorial Material	1	0.0	0	10

## Authorship

Although the focus is on the research content, it might still be relevant to show some data at the individual level in

Table 2 while then focusing on researchers' co-authorship with other researchers in the included publications identified in WoS (Figure 2).

Table 2: Author level data.

<b>AUTHOR</b>	<b>RECS</b>	<b>TLCS</b>	<b>TGCS</b>
Liu Y	31	131	470
Kumar A	23	180	366
Li Y	21	116	249
Lu WZ	21	261	692
Zhang L	21	20	224
Wang Y	20	16	83
Mlakar P	19	143	275
Perez P	17	340	629
Zhang Y	16	127	321
Li Q	15	177	337
Ma J	15	14	126
Oprea M	15	33	85
Kolehmainen M	14	519	1,088
Liu H	14	28	87
Schwartz J	14	115	391
Wang JZ	14	187	463
Li X	13	135	254
Nieto PJG	13	118	217
Wang ZY	13	76	230

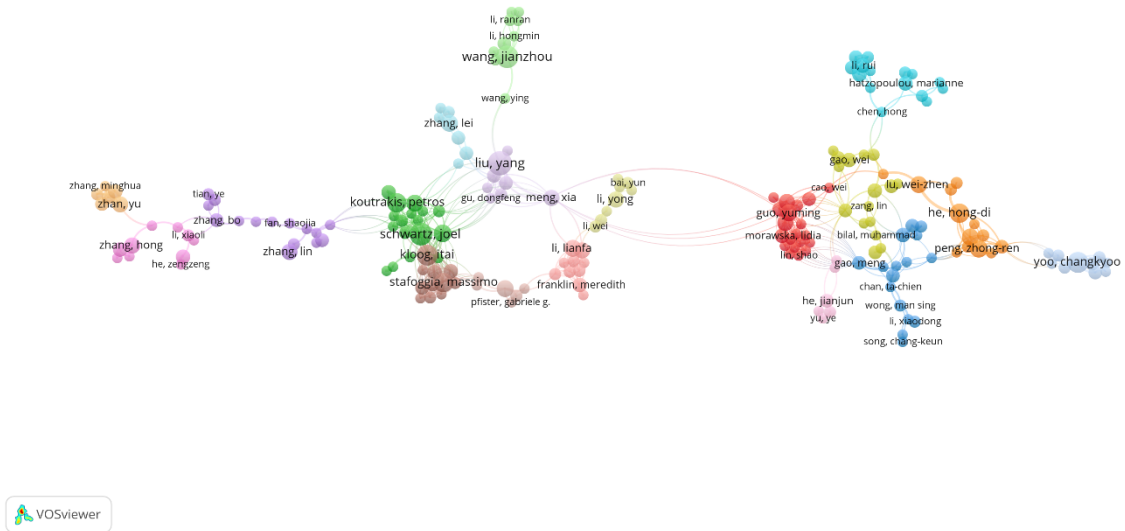


Figure 2: Co-authorship – Authors. Of 10,367 authors, 666 (321) were found  $\geq 3$  times. Visualisation: CoAuthAuth.png/pdf

## Organisation level

A substantial share of the most prolific universities in the selection consists of Asian and Middle Eastern universities (Table 3). In Figure 3, instead, we see the full breadth of collaboration with organisations with at least three authorship contributions in the data set. Note that there are two visualisations, one using clusters to differentiate potential thematic clusters and one which uses the average publication year for each institution's contribution to the data set. Visualisations do not accompany the subdivision into smaller institutional units in Table 6 and the aggregation at the country level in Table 7.

Table 3 Organisation

#	INSTITUTION	RECS	PERCENT	TLCS	TGCS
1	Chinese Acad Sci	83	2.6	315	1,333
2	Tsinghua Univ	43	1.4	180	629
3	Nanjing Univ Informat Sci & Technol	41	1.3	51	227
4	Wuhan Univ	38	1.2	109	469
5	Peking Univ	36	1.1	244	534
6	City Univ Hong Kong	34	1.1	317	1,028
7	Shanghai Jiao Tong Univ	29	0.9	83	275
8	Univ Chinese Acad Sci	28	0.9	138	280
9	Aristotle Univ Thessaloniki	27	0.9	215	540
10	Sun Yat Sen Univ	26	0.8	92	302
11	Univ Tehran	26	0.8	60	326
12	North China Elect Power Univ	25	0.8	75	237
13	Islamic Azad Univ	24	0.8	60	413
14	Lanzhou Univ	24	0.8	246	537
15	NASA	24	0.8	101	447
16	Zhejiang Univ	24	0.8	126	351
17	Emory Univ	23	0.7	36	141
18	Beijing Univ Technol	22	0.7	28	87
19	Dongbei Univ Finance & Econ	22	0.7	233	610
20	CNR	21	0.7	49	203
21	Indian Inst Technol	21	0.7	119	381

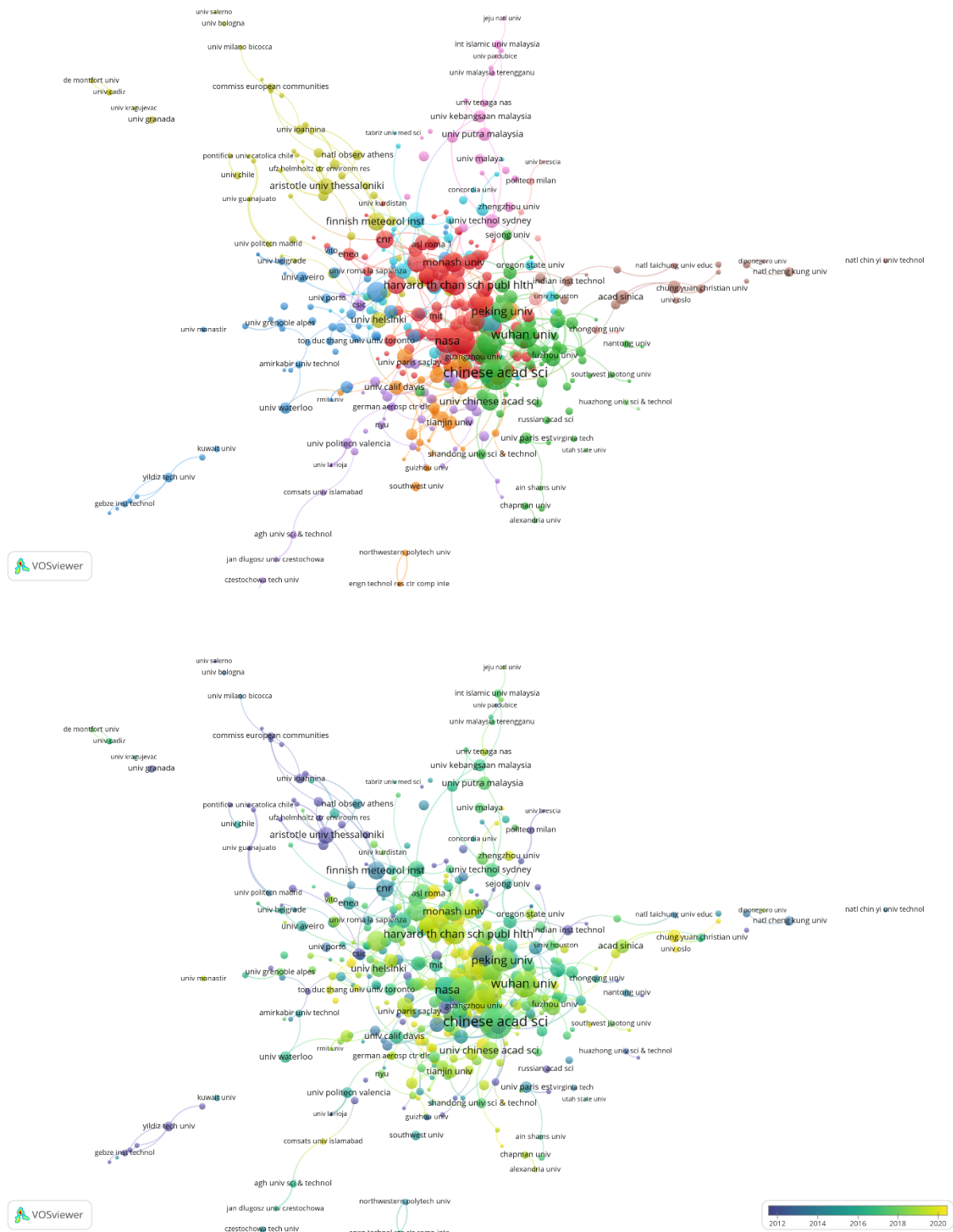


Figure 3: Co-authorship – Organisations. Of 2,952 organisations, 481 (570) were found  $\geq 3$  times. Top: Clusters, Bottom: Average Publication year. Visualisation: CoAuthOrg.png/pdf



Table 4: Institution with subdivision

INSTITUTION WITH SUBDIVISION	RECS	PERCENT	TLCS	TGCS
Dongbei Univ Finance & Econ, Sch Stat	22	0.7	233	610
Univ Chinese Acad Sci	21	0.7	138	261
City Univ Hong Kong, Dept Bldg & Construct	19	0.6	271	778
Chinese Acad Sci, Inst Geog Sci & Nat Resources Res	18	0.6	64	238
Emory Univ, Rollins Sch Publ Hlth	17	0.5	26	111
NASA, Goddard Space Flight Ctr	17	0.5	101	376
Univ Santiago Chile, Dept Fis	14	0.4	334	617
Tsinghua Univ, Sch Environm	13	0.4	79	244
Unknown	13	0.4	6	25
Aristotle Univ Thessaloniki, Dept Mech Engr	12	0.4	89	162
Beijing Univ Technol, Fac Informat Technol	12	0.4	11	44
Harvard TH Chan Sch Publ Hlth, Dept Environm Hlth	12	0.4	37	126
Univ Kuopio, Dept Environm Sci	12	0.4	449	966
Wuhan Univ, State Key Lab Informat Engr Surveying M...	12	0.4	69	230
Indian Inst Technol, Dept Civil Engr	11	0.3	70	205
Jozef Stefan Inst	11	0.3	6	55
Lanzhou Univ, Sch Math & Stat	11	0.3	167	342
Univ E Anglia, Sch Environm Sci	11	0.3	523	1,145
Univ Ioannina, Dept Phys	11	0.3	144	358
Wuhan Univ, Sch Resource & Environm Sci	11	0.3	54	182

Table 5: Country

#	COUNTRY	RECS	PERCENT	TLCS	TGCS
1	Peoples R China	899	28.4	2,865	10,872
2	USA	532	16.8	1,569	8,546
3	Italy	193	6.1	865	3,400
4	UK	193	6.1	815	3,714
5	India	179	5.7	509	1,560
6	Spain	165	5.2	430	2,337
7	Iran	126	4.0	236	1,300
8	South Korea	106	3.4	139	819
9	Germany	104	3.3	252	1,365
10	Taiwan	101	3.2	232	1,070
11	Australia	95	3.0	257	1,417
12	Turkey	90	2.8	386	956
13	Canada	89	2.8	228	1,355
14	Greece	87	2.8	733	2,302
15	France	83	2.6	269	1,381
16	Poland	77	2.4	93	625
17	Malaysia	71	2.2	81	602
18	Brazil	59	1.9	118	613
19	Unknown	51	1.6	227	645
20	Romania	46	1.5	43	238
30	Sweden	29	0.9	41	248

## Journal level

When viewed as a top list of journals, we find a broad range of publication outlets in Table 6. Presenting the distribution as a network map of journals as in Figure 4, we can find patterns in the results within the total of 1,409 different publication outlets where the research was published. The bibliographic coupling algorithm is used here, meaning that two journals are closely connected based on the overlap of reference lists in their respective published articles. The node sizes are based on the relative number of published articles within each journal. The first one shows topics as different colours based on the layout algorithm used, while the second graph is colour coded based on the average publication year for each journal within the set.

Table 6: Journal sources

JOURNAL	RECS	PERCENT	TLCS	TGCS
ATMOSPHERIC ENVIRONMENT	147	4.6	2,593	5,999
SCIENCE OF THE TOTAL ENVIRONMENT	83	2.6	845	2,097
IEEE ACCESS	55	1.7	100	284
ENVIRONMENTAL POLLUTION	54	1.7	456	1,269
ATMOSPHERIC POLLUTION RESEARCH	48	1.5	415	791
AIR QUALITY ATMOSPHERE AND HEALTH	42	1.3	153	323
JOURNAL OF CLEANER PRODUCTION	42	1.3	82	415
SENSORS	39	1.2	2	336
BUILDING AND ENVIRONMENT	37	1.2	128	1,018
INTERNATIONAL JOURNAL OF ENVIRONMENTAL RESEARCH AND PUBLIC HEALTH	34	1.1	31	283
ATMOSPHERE	32	1.0	5	161
ENVIRONMENTAL SCIENCE AND POLLUTION RESEARCH	32	1.0	168	442
SUSTAINABILITY	32	1.0	0	137
JOURNAL OF THE AIR & WASTE MANAGEMENT ASSOCIATION	30	0.9	385	803
APPLIED SCIENCES-BASEL	29	0.9	0	115
ENVIRONMENTAL SCIENCE & TECHNOLOGY	29	0.9	192	691
ENVIRONMENTAL MODELLING & SOFTWARE	28	0.9	422	1,669
REMOTE SENSING	26	0.8	0	135
ENVIRONMENT INTERNATIONAL	24	0.8	157	561
CHEMOSPHERE	23	0.7	182	853

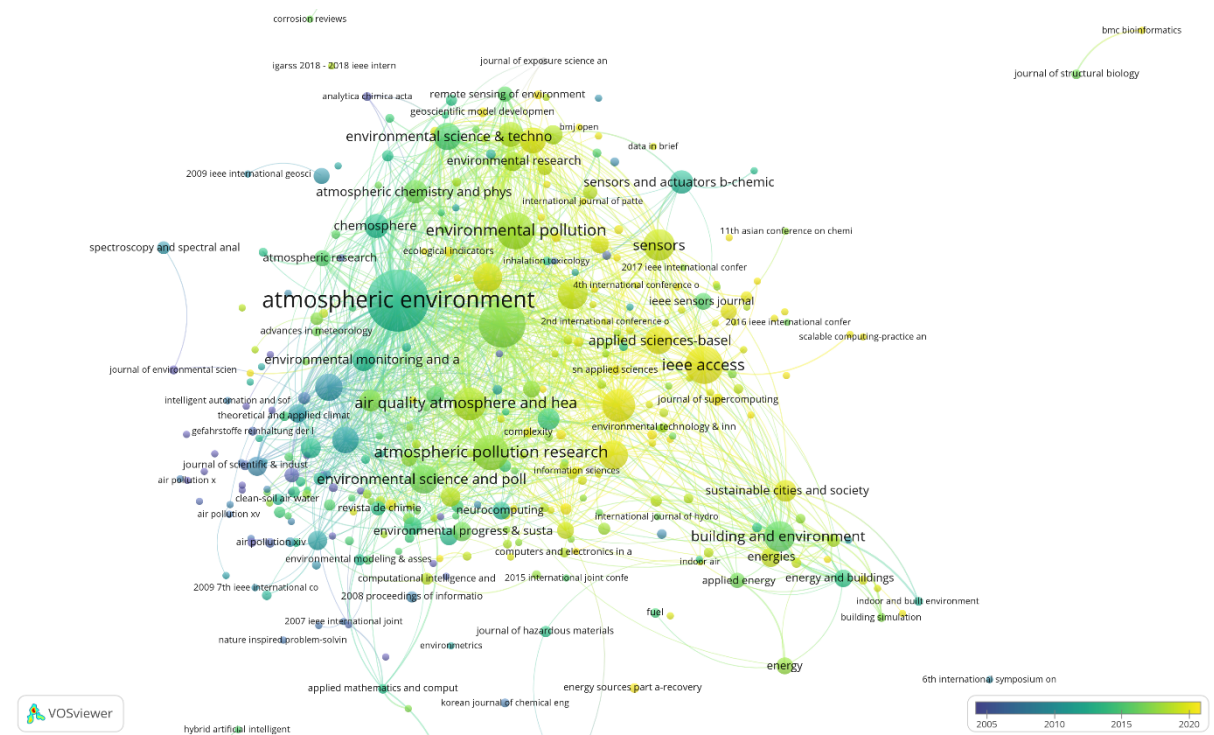
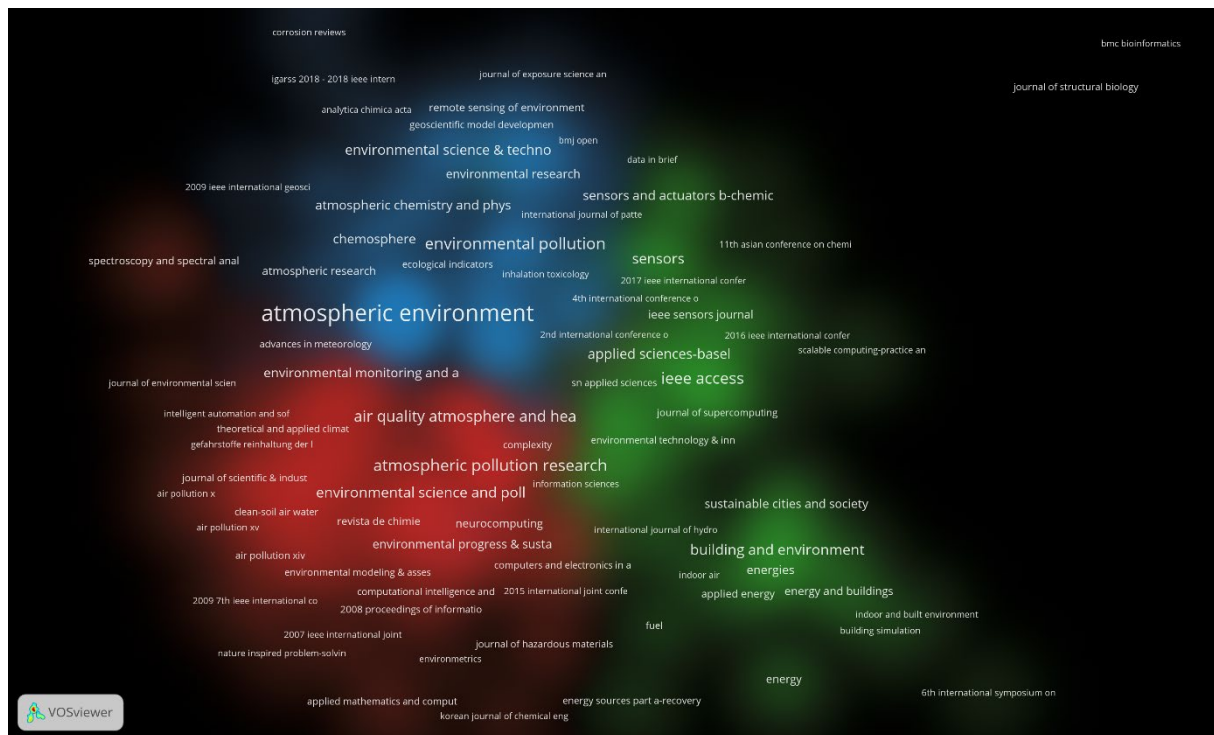


Figure 4: Bibliographic coupling – Sources. Of 1,409 sources, 334 (330) had  $\geq 2$  documents. Top: Density view, Bottom: Nodes clustered based on average publication year. Visualisation: BiblCoupSO.png/pdf; BiblCoupSOPY.png/pdf

## Co-citation analysis

Lastly, we show the most cited documents (Figure 5) and all source outlets (Figure 6) that the researchers in the dataset cite within the publications. The Co-citation network shows the “intellectual basis” of the collective (Persson, 1994). Highly cited articles and source outlets contain the most frequently used research among the collected data. The articles cluster around specific topics, as can be seen through the titles of the sources. Both the green cluster on top and the red one to the right mainly consists of articles in the journal with the shortened title *atmos environ*. However, since these clusters are separated, these articles seem to cover different topics. Specifically, it can be seen that there is a temporal pattern, where the red cluster consists of articles published in the first decade of the 21<sup>st</sup> century.

In contrast, the green one consists of articles from the second decade. The purple cluster at the bottom consists of articles focusing on sensors, while the blue to the right seems to consist of interdisciplinary literature in engineering and environmental science. The yellow cluster, which divides the cited literature into two parts, seems to consist of a more fundamental kind of machine learning literature. This research is applied in the research found in the more distant clusters. Some of the details are lost when the data is aggregated at the journal level. It is because one journal stands out so much in terms of the numbers of citations. However, we find that a significant share of the cited literature is published in traditional disciplines, whether geophysics, remote sensing or the built environment.

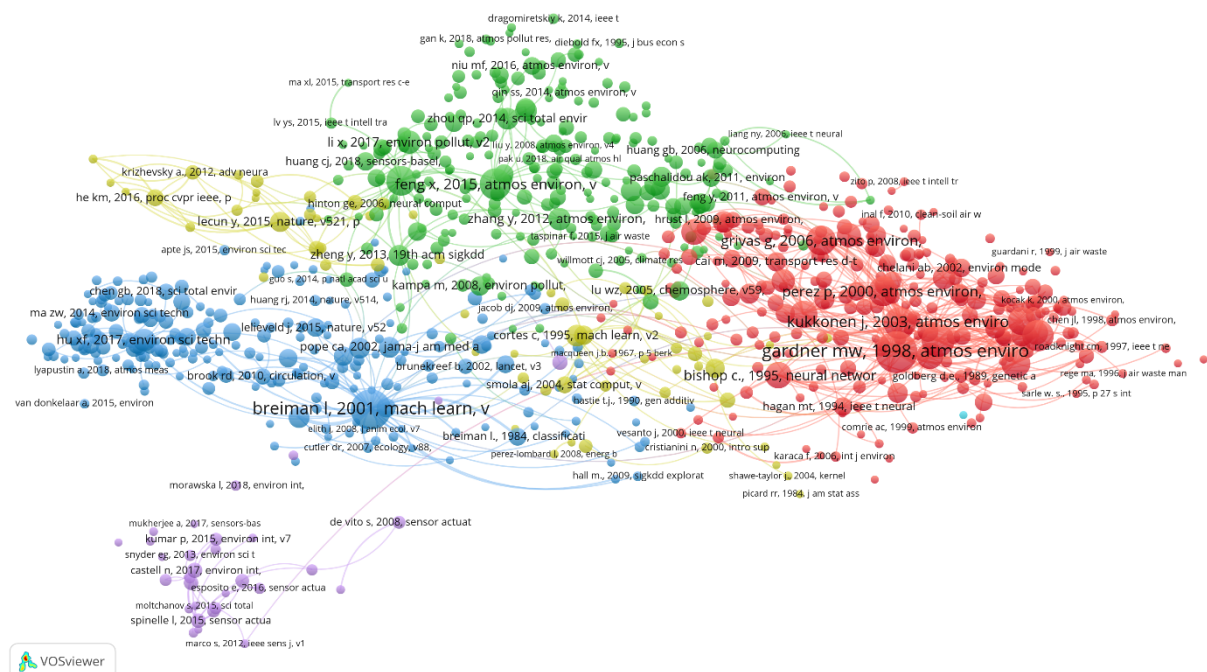


Figure 5: Co-citation – Documents. Of 81,062 sources, 818 had  $\geq 10$  citations. Visualization: CoCitDO.png/pdf

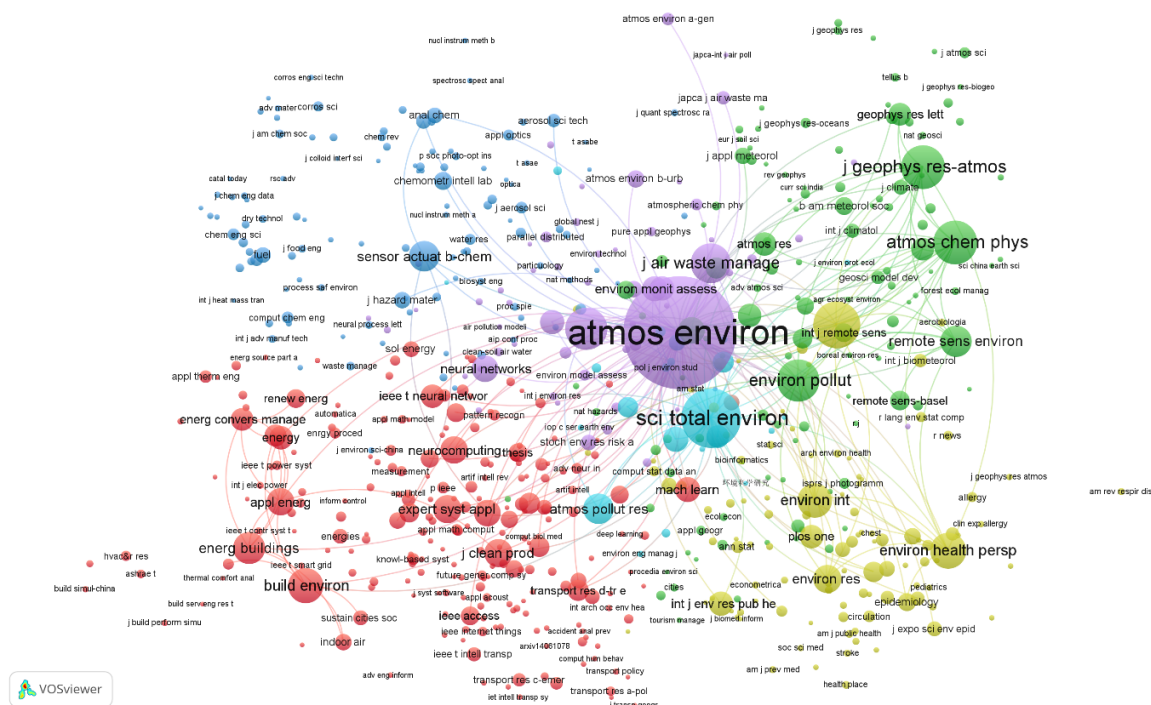


Figure 6: Co-citation – Sources. Of 25,644 sources, 644 had  $\geq 20$  citations. Visualization: CoCitSO.png/pdf

## Keywords and co-word analysis

Lastly, we employ two techniques to identify thematic information about the research content covered by the articles identified in our searches. We select the most frequently used author keywords chosen for the publications (Figure 7, Figure 8). Another technique, co-word analysis, extract nouns and so-called noun phrases, using computer linguistic methods to identify phrases of text could sometimes elucidate more specific information (Figure 9).

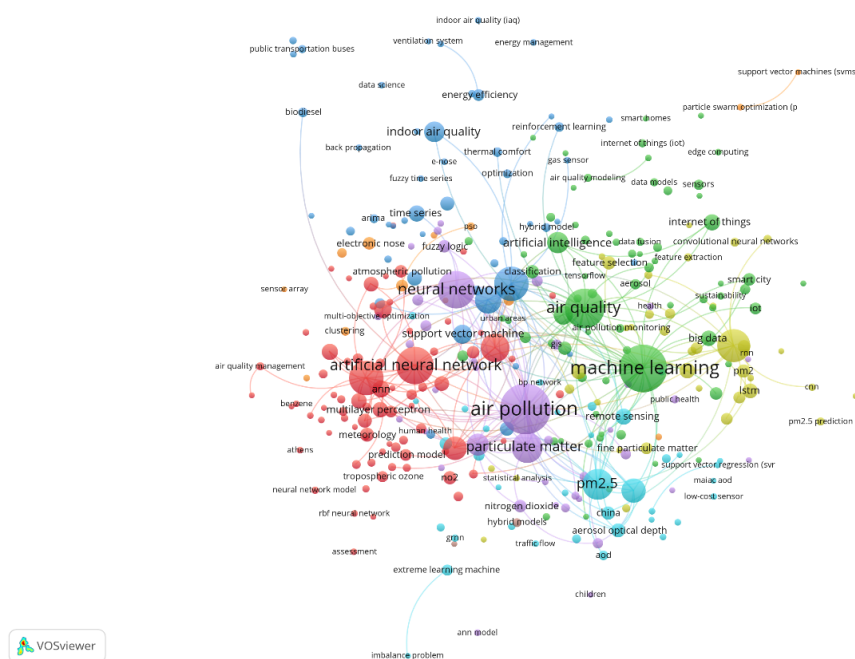


Figure 7. Author keywords. Of 6,517 keywords, 323 (323) were found  $\geq 5$  times. Visualisation: keywords.png/pdf



We add some details of the map to show that different machine learning algorithms seem to be related to specific research areas. Colours show which topic cluster each specific keyword belongs to:

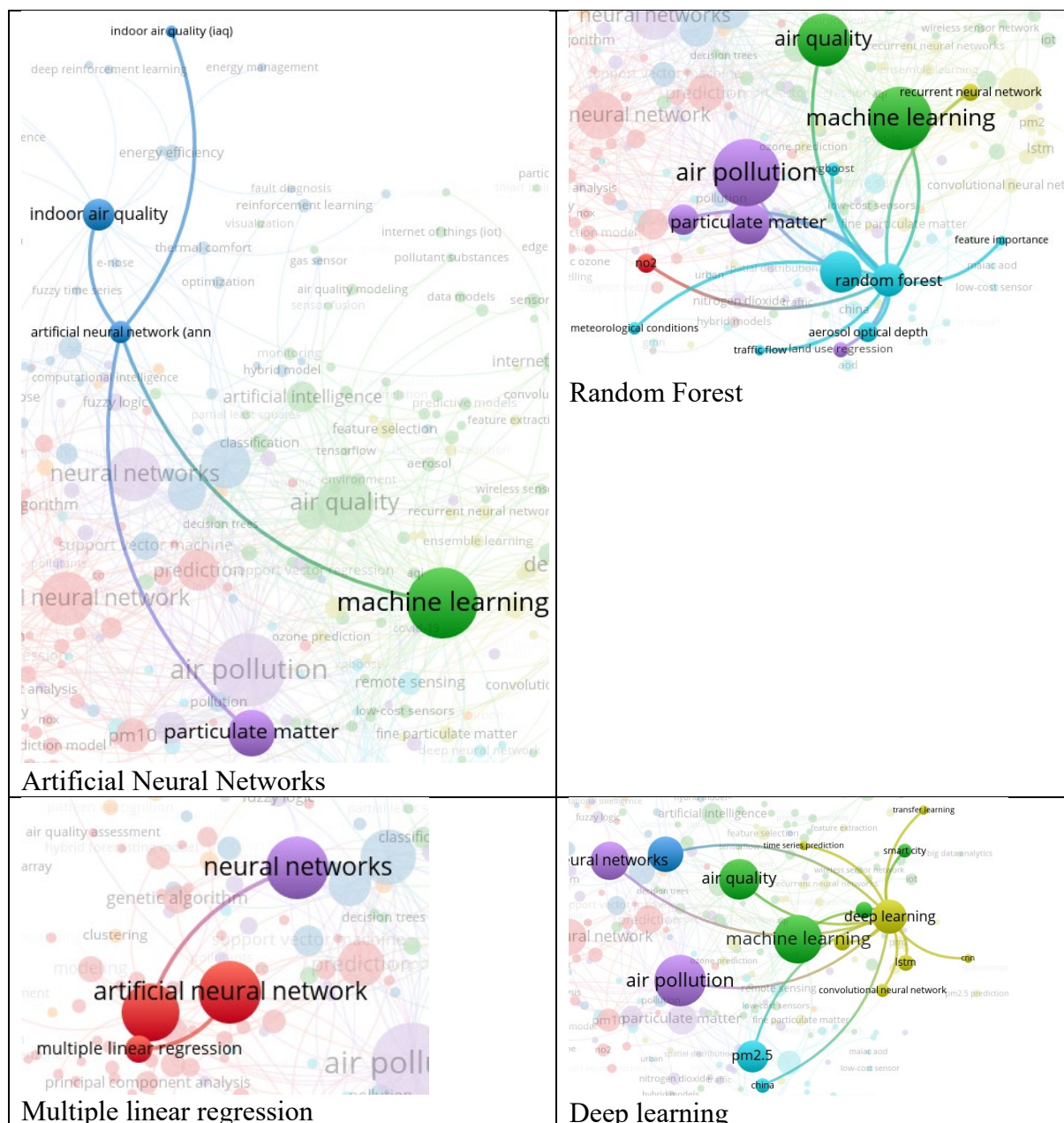


Figure 8(A-D): Distinct machine learning terms within the keyword co-occurrence map.

In the co-word analysis of relevant terms identified in the titles and abstracts of the retrieved documents, we identify terms and phrases relevant to different domains of analysis and specific topics. In a sense, it is possible to identify topics relating to research about air quality issues and AI methodology. The following description is tentative but shows one way of ascribing a “story” to the terms and phrases found in the co-word map. In the blue cluster, we find terms relating to forecasting and prediction technology. The green cluster covers particle matter, while the yellow cluster at the bottom depicts terms related to remote sensing approaches and epidemiological studies. The red cluster holds terms relating to instruments and identifies air quality issues in more general terms, while the purple cluster relates to indoor quality.

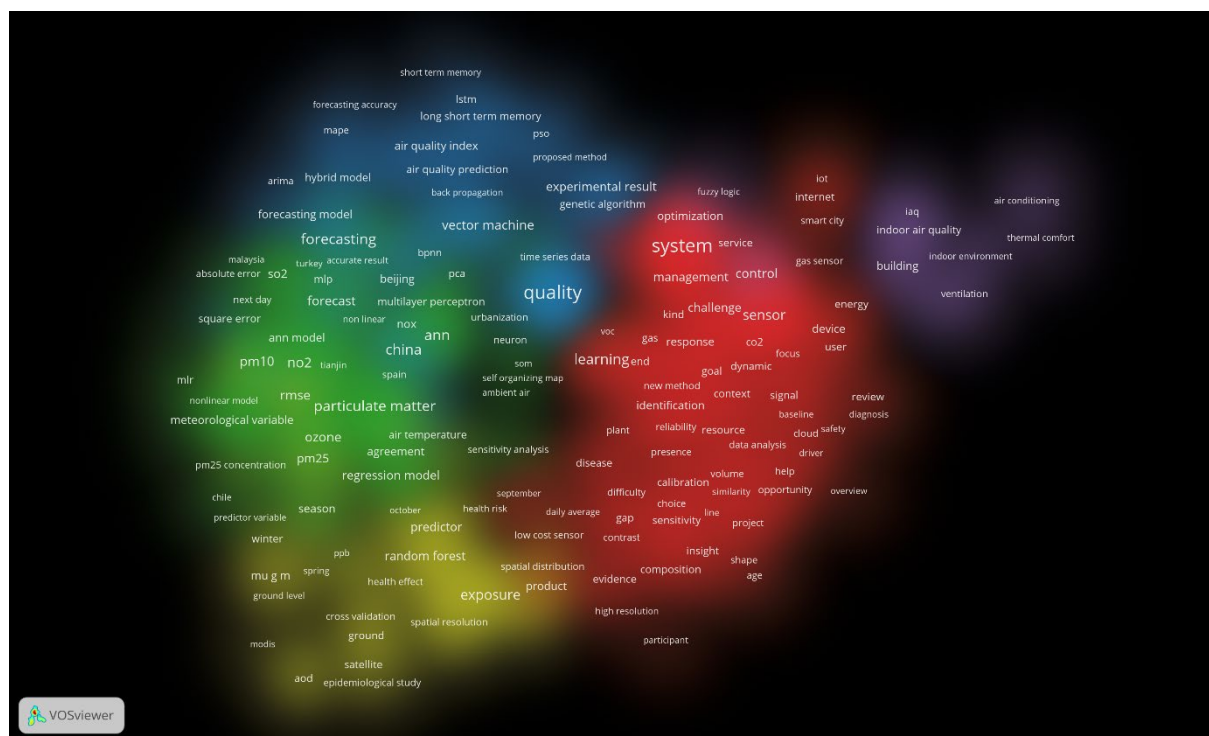


Figure 9: Co-word analysis. Of 64,048 noun phrases, 843 (506, TF-IDF=60%) where found  $\geq 20$  times. Visualization: cowords.png/pdf

## Scientometric report, Traffic safety research

*(All data were downloaded from Web of Science Core Collection on 2021-02-19)*

**Timespan:** open time window. **Indexes:** SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI.

### **Search string:**

TS= (traffic OR transportation\* OR road\*)

AND

TS= (assessm\* OR safety OR crash\* OR accident\* or collision\* )

AND

TS=(("machine learning" OR "deep learning" OR "artific\* intel\*" OR "neural network\*" OR "support vector machine\*" OR "reinforcement learning" OR "random forest\*"))

**Identified documents:** 4.171

## Results

### Document type and publication year

For the publications identified in Web of Science, we show the number of documents published on a yearly basis (Figure 10). We find the number of articles per year is increasing, especially since 2014. Again, since no strict criteria for limiting the inclusion were performed, the citation database included some documents published in 2021 and a few with an “unknown” date. The latter are generally preprints that have yet to receive a publishing date.

Two additional columns are shown. The last one, GCS, stands for Global Citation Score and corresponds to the number of citations the documents at a specific year has received to the time of extraction of data. The middle one, LCS, stand for “Local Citation Score” and shows the number of citations the publications of a particular year has received within the data set.



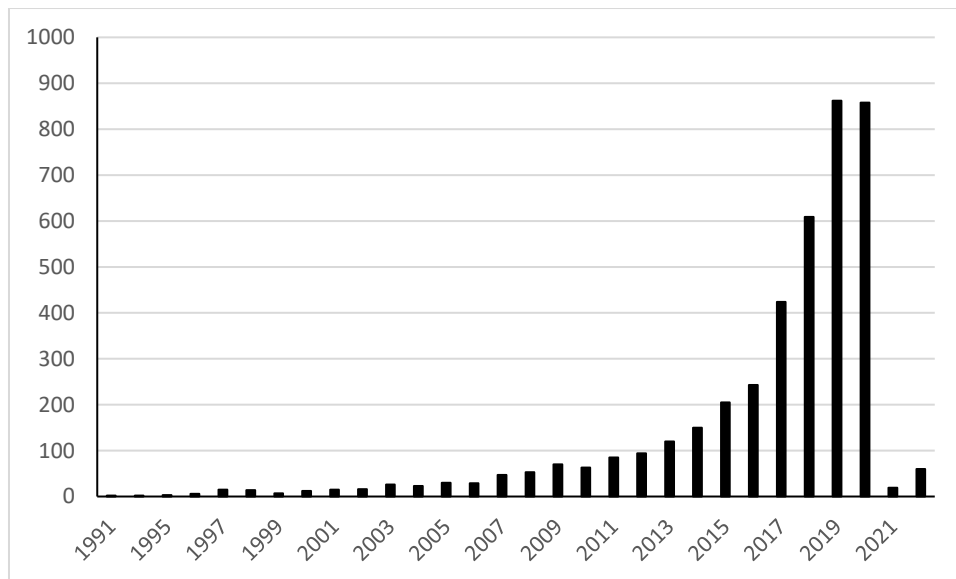


Figure 10. The number of yearly publications since 1991. The last column shows “unknown year, which primarily relates to ahead-of-print publications.

Most documents are of the category (peer reviewed) “article”. A pretty large share of the documents is proceedings papers, which is quite common in engineering sciences. There is also a small number, but in terms of citations, quite significant publications of the review kind. Other document types were relatively few. Although some of these are not peer reviewed, it was deemed unnecessary to remove these publications since they might still be relevant for information purposes. As with self-citations, as will be seen later, when using bibliographic data for exploratory tasks, less strict inclusion criteria is often warranted.

Table 7: Document type

DOCUMENT TYPE	RECS	PERCENT	TLCS	TGCS
Article	2,377	57.1	3,533	37,683
Proceedings Paper	1,576	37.9	470	4,852
Review	81	1.9	68	1,920
Article; Proceedings Paper	54	1.3	142	851
Article; Early Access	53	1.3	0	22
Review; Early Access	7	0.2	0	6
Editorial Material	4	0.1	0	3
Article; Data Paper	2	0.0	0	0
Letter	2	0.0	0	0
Meeting Abstract	2	0.0	0	0
Article; Book Chapter	1	0.0	3	78
Proceedings Paper; Retracted Publication	1	0.0	0	2
Review; Book Chapter	1	0.0	0	12
Software Review	1	0.0	0	11

## Authorship

Although the focus is on the research content, it might still be relevant to show some data at the individual level in

Table 2 while then focusing on the co-authorship of researchers with other researchers in the included publications identified in WoS (Figure 2).

Table 8: Author level data.

AUTHOR	RECS	TLCS	TGCS
Pradhan B	42	398	3,498
Abdel-Aty M	38	253	1,047
Wang C	33	29	248
Bui DT	29	135	1,875
Liu Y	26	3	96
Wang Y	26	24	170
Pourghasemi HR	25	251	2,151
Li J	23	6	146
Chen W	22	134	1,245
Li Y	18	5	47
Wang H	17	15	127
Hong HY	16	109	984
Liu J	16	3	87
Liu P	16	107	372
Wang W	16	89	260
Zhang Y	16	6	119

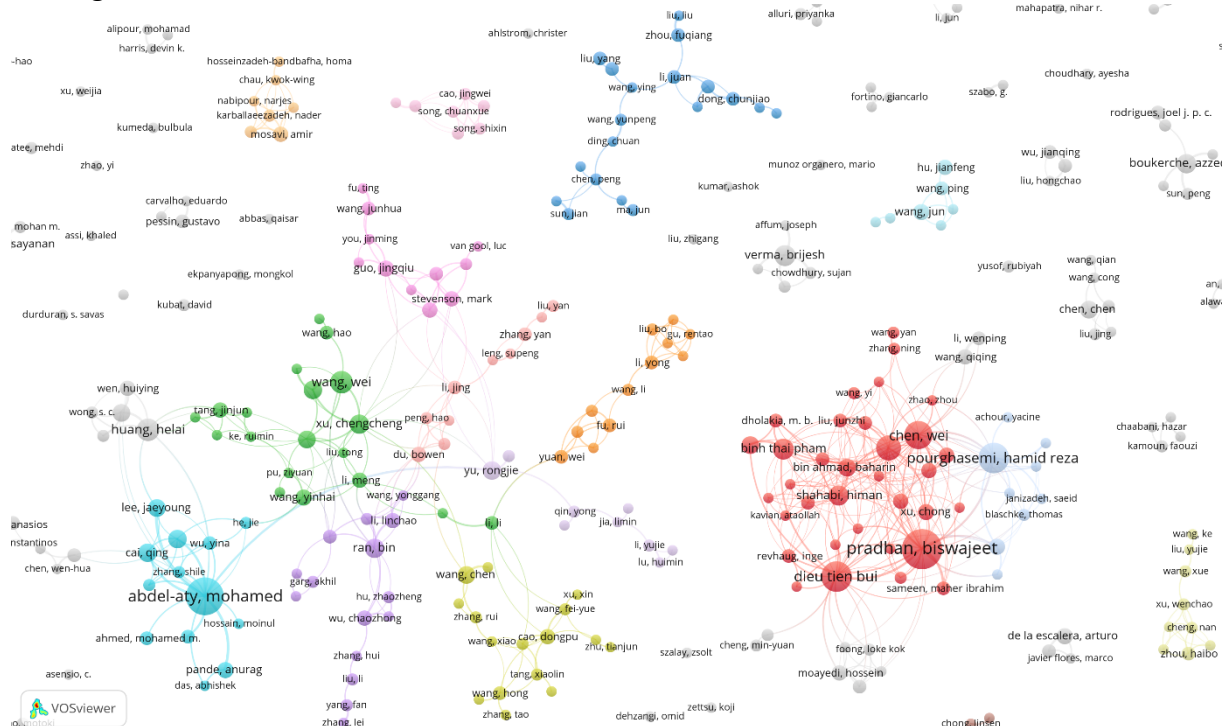


Figure 11: Co-authorship – Authors. Of 13,069 authors, 525 (321) were found  $\geq 3$  times. Visualization: CoAuthAuth.png/pdf

## Organisation level

A large share of the most prolific universities in the selection consists of Asian and Middle Eastern universities. However, as opposed to the Air Quality research set, several US, British and Australian universities also occur. (Table 9). In Figure 12Figure 3, instead, we see the full breadth of collaboration with organisations with at least three authorship contributions in the data set. Note that there are two visualisations, one using clusters to differentiate potential thematic clusters and one which uses the average publication year for each institution's contribution to the data set. The distribution is very heterogeneous, and there are no obvious co-authorship patterns at any level of aggregation.

Regarding the average year of publication, there does not seem to develop new clusters of actors over time. Instead, the average year of publication seems to be in the middle of the range, indicating that the organisations have distributed output coverage over the years. The subdivision into smaller institutional units in Table 10 and the aggregation at the country level in Table 11 are not accompanied by visualisations.

Table 9: Organisation

INSTITUTION	RECS	PERCENT	TLCS	TGCS
Southeast Univ	69	1.7	160	573
Chinese Acad Sci	61	1.5	113	1,840
Beijing Jiaotong Univ	60	1.4	31	342
Tongji Univ	54	1.3	106	573
Tsinghua Univ	54	1.3	74	559
Univ Cent Florida	53	1.3	413	1,614
Beihang Univ	47	1.1	57	517
Changan Univ	47	1.1	33	243
Univ Technol Sydney	36	0.9	50	695
Univ Waterloo	35	0.8	22	543
Jilin Univ	32	0.8	5	231
Wuhan Univ Technol	32	0.8	13	111
Islamic Azad Univ	30	0.7	73	751
Univ Michigan	29	0.7	32	432
Duy Tan Univ	26	0.6	10	274
Southwest Jiaotong Univ	26	0.6	57	399
Hong Kong Polytech Univ	25	0.6	11	196
MIT	25	0.6	40	461
Nanyang Technol Univ	25	0.6	8	116
Univ Illinois	25	0.6	14	225
Univ Putra Malaysia	25	0.6	307	2,528
Univ Tehran	25	0.6	26	717

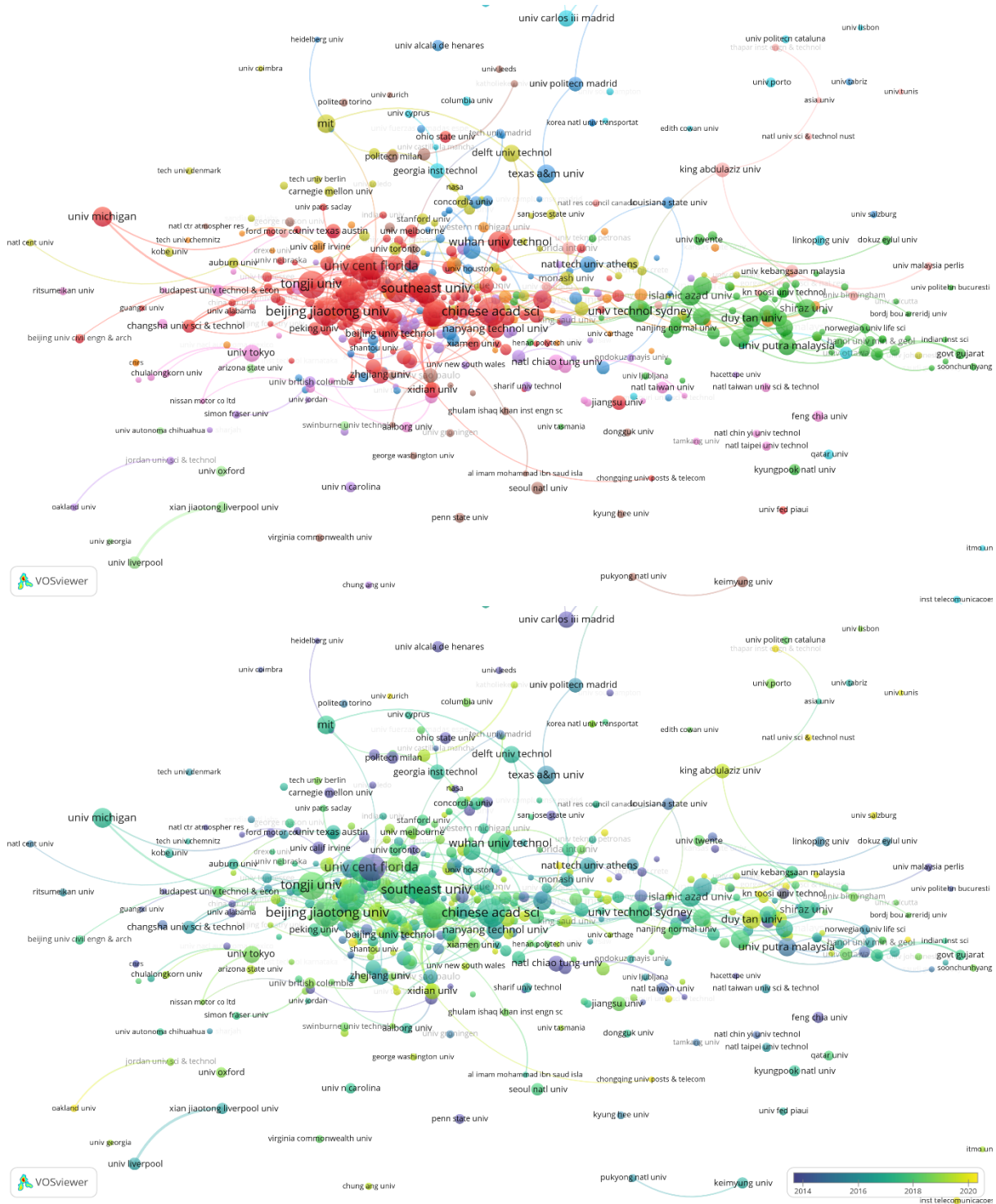


Figure 12 Co-authorship – Organisations. Of 3,431 organisations, 643 (564) were found  $\geq 3$  times. Top: Clusters, Bottom: Average Publication year. Visualisation: CoAuthOrg.png/pdf

Table 10: Institution with subdivision

INSTITUTION WITH SUBDIVISION	RECS	PRCN	TLCS	TGCS
Univ Cent Florida, Dept Civil Environm & Construct Engr	37	0.9	234	835
Southeast Univ, Sch Transportat	28	0.7	132	366
Duy Tan Univ, Inst Res & Dev	24	0.6	10	244
Unknown	19	0.5	9	112
Xian Univ Sci & Technol, Coll Geol & Environm	18	0.4	109	1,034
Beijing Jiaotong Univ, Sch Traff & Transportat	17	0.4	10	63
Sejong Univ, Dept Energy & Mineral Resources Engr	15	0.4	44	510
Univ Putra Malaysia, Fac Engr	15	0.4	232	1,671
South China Univ Technol, Sch Civil Engr & Transportat	14	0.3	18	168
Cent South Univ, Sch Traff & Transportat Engr	13	0.3	47	199
Shiraz Univ, Coll Agr	13	0.3	45	589
Beijing Jiaotong Univ, State Key Lab Rail Traff Control & Safety	12	0.3	3	163
Tongji Univ, Key Lab Rd & Traff Engr	12	0.3	26	94
Tsinghua Univ, State Key Lab Automot Safety & Energy	12	0.3	4	70
Univ Chinese Acad Sci	12	0.3	0	23
Southeast Univ, Jiangsu Key Lab Urban ITS	11	0.3	26	122
Univ Technol Sydney, Fac Engr & IT	11	0.3	41	373
Chinese Acad Sci, Inst Automat	10	0.2	43	1,002
Hong Kong Polytech Univ, Dept Comp	10	0.2	5	29
Jilin Univ, State Key Lab Automot Simulat & Control	10	0.2	0	45
Ton Duc Thang Univ, Dept Management Sci & Technol Dev	10	0.2	10	147
Univ Cent Florida, Dept Civil & Environm Engr	10	0.2	179	570
Univ Kurdistan, Fac Nat Resources	10	0.2	51	699
Univ Washington, Dept Civil & Environm Engr	10	0.2	37	138

Table 11: Country level

COUNTRY	RECS	PERCENT	TLCS	TGCS
Peoples R China	1,194	28.7	1,132	10,806
USA	898	21.6	1,361	12,565
India	256	6.2	189	2,027
South Korea	197	4.7	270	3,040
Canada	166	4.0	73	1,513
Australia	161	3.9	105	1,800
UK	159	3.8	133	3,853
Germany	158	3.8	114	1,654
Japan	158	3.8	149	1,443
Iran	150	3.6	445	4,564
Spain	138	3.3	78	1,371
Taiwan	120	2.9	78	1,202
Italy	119	2.9	117	1,599
France	104	2.5	56	1,125
Turkey	101	2.4	302	1,927
Malaysia	97	2.3	429	3,825
Saudi Arabia	66	1.6	105	588
Netherlands	65	1.6	43	1,419
Brazil	61	1.5	15	308
Singapore	58	1.4	66	652

## Journal level

We find a broad range of publication outlets at the journal level in Table 6, while the breadth is drastically extended when we see all 2,023 journals where the research was published in Figure 4. The bibliographic coupling algorithm is used here, meaning that two journals are closely connected based on the overlap of reference lists in their respective published articles. There are three distinct clusters, containing research in transportation and accident analyses (top), environmental sciences and the geosciences (right), and intelligent transport systems research sensors and engineering, found to a large degree in conference publications (bottom). A less distinct cluster, focusing on infrastructures, structures and the built environment (yellow), is also found.

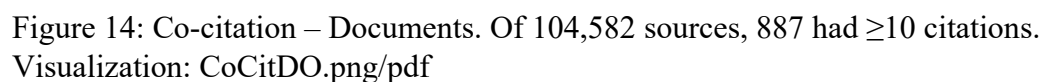
Table 12: Journal sources

JOURNAL	RECS	PERCENT	TLCS	TGCS
IEEE ACCESS	152	3.7	71	434
IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS	96	2.3	272	3,575
SENSORS	95	2.3	62	925
ACCIDENT ANALYSIS AND PREVENTION	92	2.2	788	2396
TRANSPORTATION RESEARCH RECORD	71	1.7	91	525
APPLIED SCIENCES-BASEL	49	1.2	0	249
TRANSPORTATION RESEARCH PART C-EMERGING TECHNOLOGIES	47	1.1	153	1,231
IET INTELLIGENT TRANSPORT SYSTEMS	40	1.0	57	287
JOURNAL OF ADVANCED TRANSPORTATION	34	0.8	8	166
EXPERT SYSTEMS WITH APPLICATIONS	32	0.8	88	1,085
SUSTAINABILITY	30	0.7	0	71
REMOTE SENSING	29	0.7	0	204
ENVIRONMENTAL EARTH SCIENCES	25	0.6	62	1,059
IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY	24	0.6	36	575
NEURAL COMPUTING & APPLICATIONS	24	0.6	21	212
2019 IEEE INTELLIGENT TRANSPORTATION SYSTEMS CONFERENCE (ITSC)	23	0.6	1	6
2018 21ST INTERNATIONAL CONFERENCE ON INTELLIGENT TRANSPORTATION SYSTEMS (ITSC)	22	0.5	12	59
INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS	20	0.5	2	11
COMPUTER-AIDED CIVIL AND INFRASTRUCTURE ENGINEERING	19	0.5	38	909
ELECTRONICS	18	0.4	0	37
MATHEMATICAL PROBLEMS IN ENGINEERING	18	0.4	0	254
SAFETY SCIENCE	17	0.4	39	299
INTERNATIONAL JOURNAL OF ENVIRONMENTAL RESEARCH AND PUBLIC HEALTH	15	0.4	0	76
ISPRS INTERNATIONAL JOURNAL OF GEO-INFORMATION	15	0.4	0	79
MULTIMEDIA TOOLS AND APPLICATIONS	15	0.4	6	65





When aggregated at the journal level, we find the same issue as in the Air quality research report, that some of the details are lost since one journal stands out so much in terms of numbers of citations. However, we find that a significant share of the cited literature is published in traditional disciplines, whether geophysics, remote sensing or the built environment.





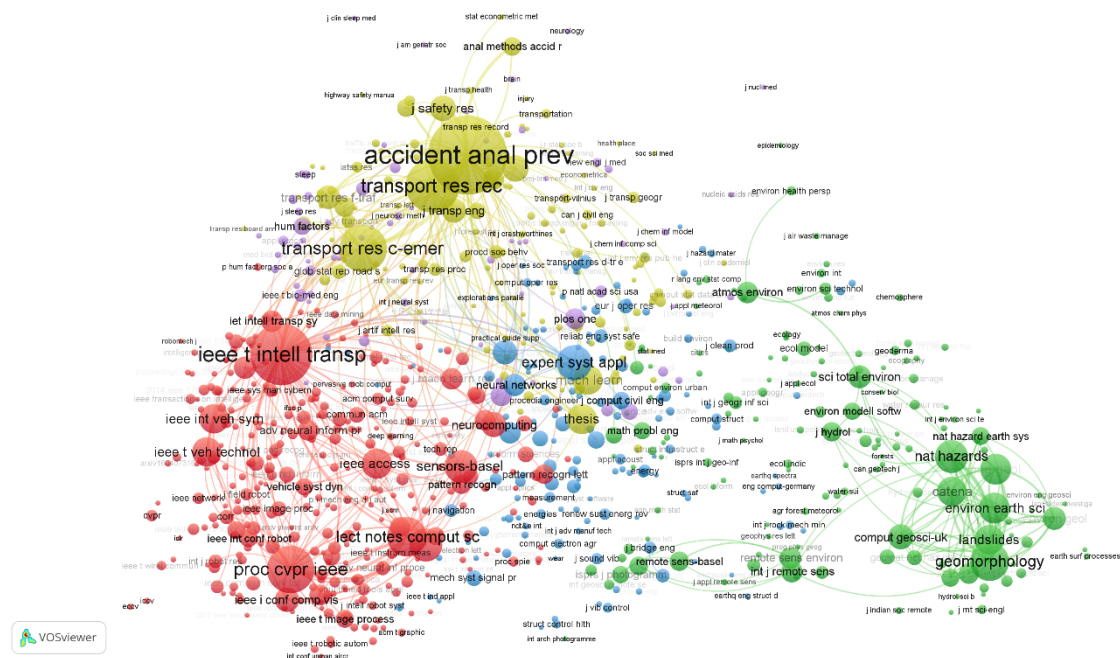


Figure 15: Co-citation – Sources. Of 38,867 sources, 806 had  $\geq 20$  citations. Visualization: CoCitSO.png/pdf

## Keywords and co-word analysis

Lastly, we employ two techniques to identify thematic information about the research content covered by the articles identified in our searches. We select the most frequently used author keywords chosen for the publications (Figure 16, Figure 17). Another technique, co-word analysis, extracts nouns and so-called noun phrases, using computer linguistic methods to identify phrases of text could sometimes elucidate more specific information (Figure 18).

In the keyword analysis, specific machine learning technologies dominate the respective clusters. ‘Deep learning’ and ‘convolutional neural networks’ are associated with image analysis and computer vision in the blue cluster. In the green cluster, the generic term artificial intelligence and ‘reinforcement learning’, a much more specific term, are found in conjunction with autonomous vehicles. The red cluster, dominated by machine learning, is quite generic, although driver behaviour and physiological factors, including EEG, are connected to the support vector machines at the top. Lastly, although this is a cursory analysis, it seems like geographic information systems and land use concepts are found further away from some of the buzz words in machine learning. The most distinct concept is the more classic machine learning algorithm: ‘logistic regression’. Possibly, we have yet to see faster developments within this area in the future?

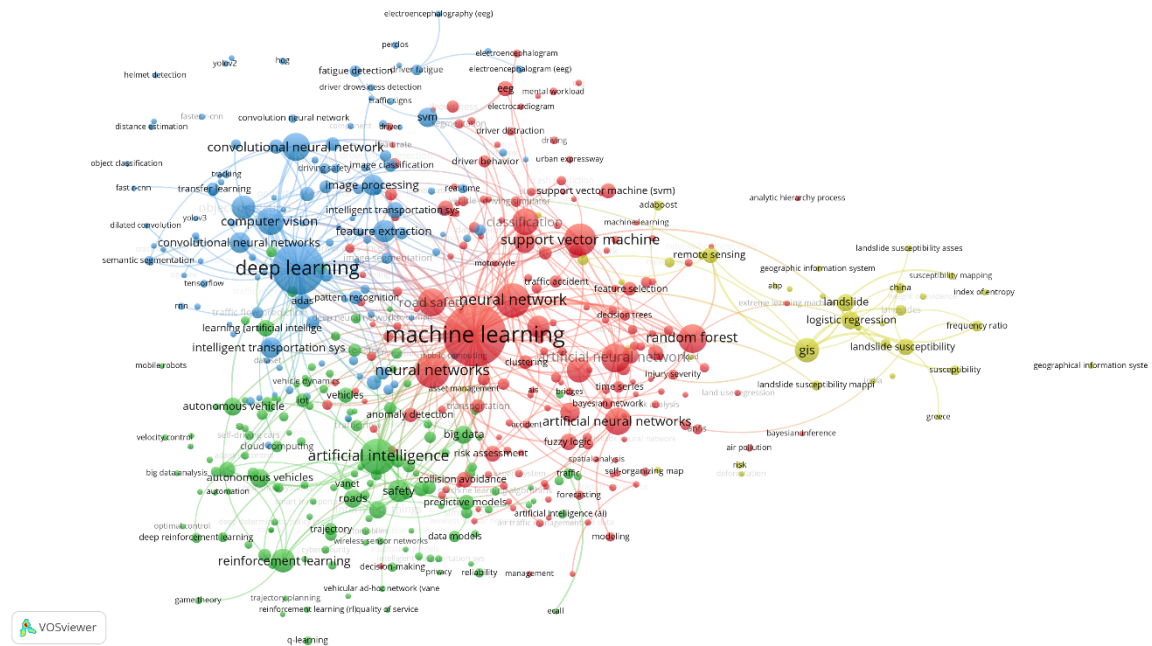
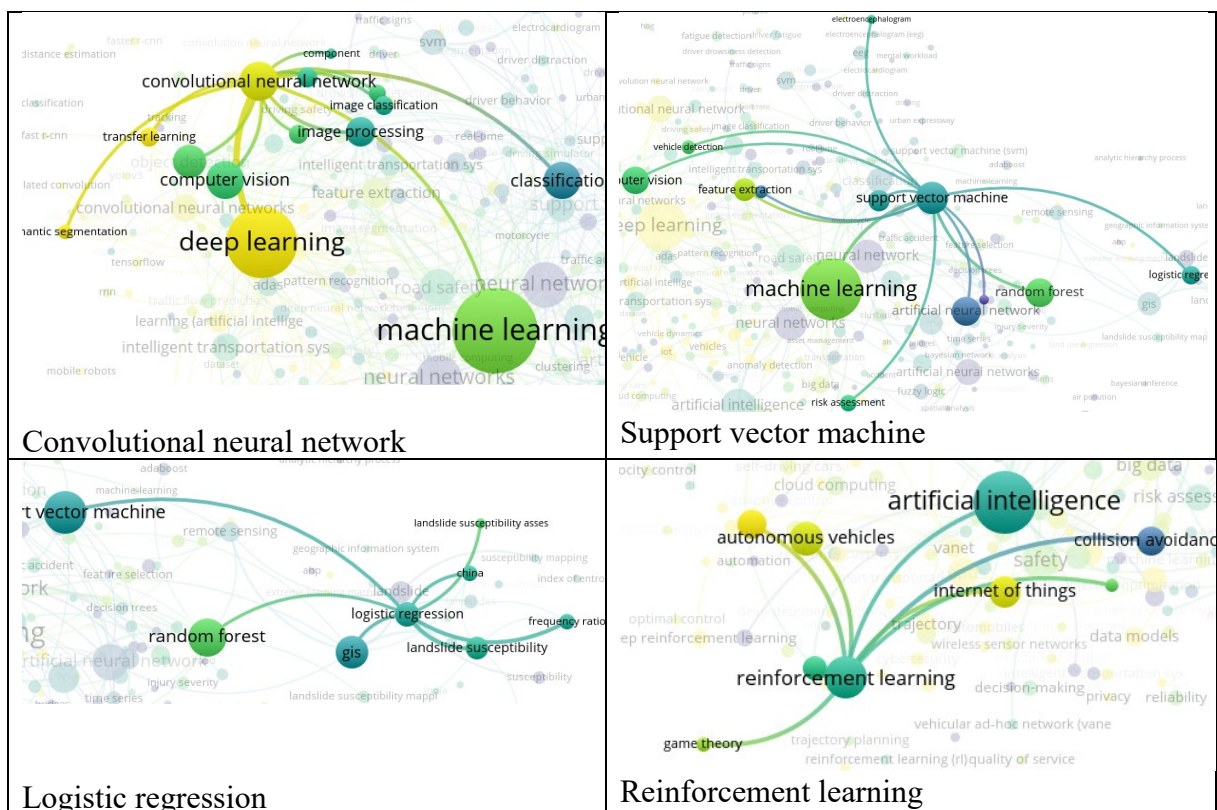


Figure 16. Author keywords. Of 9,979 keywords, 474 (473) were found  $\geq 5$  times. Visualisation: keywords.png/pdf

Below we add some details of the keyword map to show that different machine learning algorithms seem to be related to specific research areas, as noted above. Instead of Topic clustering, node colours show the average year of publication for articles using the specific keyword. Blue is earlier, yellow is more recent:



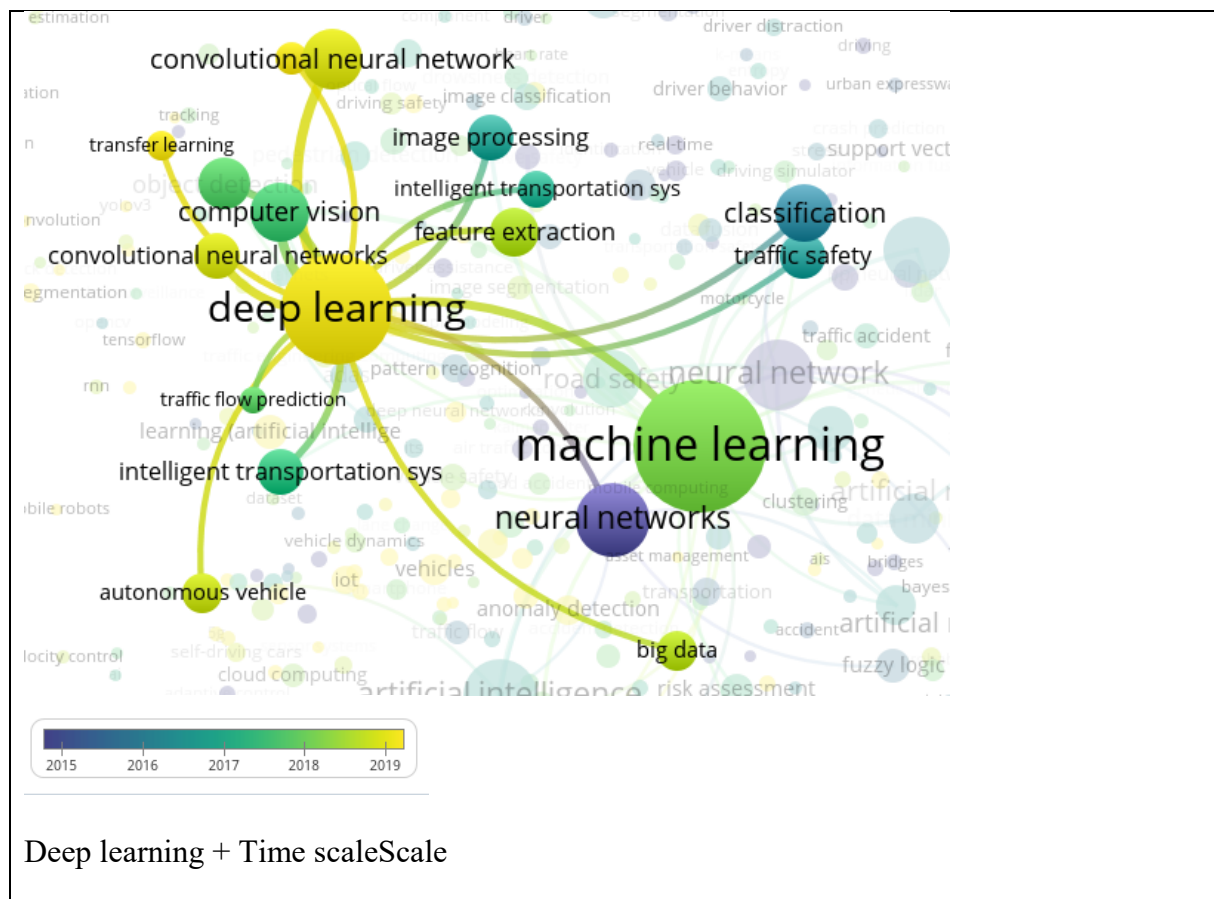


Figure 17(A-E): Distinct machine learning terms within the keyword co-occurrence map.

This study's last visualisation is based on co-word analysis of relevant terms identified in the retrieved documents' titles and abstracts. Here we identify terms and phrases relevant to different domains of analysis and specific topics. In a sense, it is possible to identify topics relating to research about air quality issues and AI methodology. The following description is tentative but shows one way of ascribing a "story" to the terms and phrases found in the co-word map. Here, four clusters are distinguished. The blue cluster contains phrases related to the driver's activity, including different detection systems, vision and psychological features like drowsiness. Instead, the green cluster focuses on the vehicle, autonomous systems and the technology of simulation and operation of the vehicle, and analyses of traffic flow, congestions, et cetera. The yellow cluster relates to features such as crashes and risk factors. In contrast, the green clusters cover the geographical settings of roads, GIS and mapping, and on the far right, the geological and geomorphological conditions.





## A scientosemantic approach to bibliographic data

This section describes an in-depth exploratory use of the bibliographic data retrieved from searches performed in Web of Science. We label this a *scientosemantic* approach to bibliographic data since we use probabilistic term weighting, computational linguistics, and machine learning to explore the content of titles and abstracts in bibliographic records. This approach complements the scientometric analyses described in the previous chapters by using results from these to investigate the content of the articles, here based on the text found in the abstracts of the articles.

To this end, we have developed tools for question answering and information searching within the retrieved data. These tools help the client identify specific methods and techniques used to measure air quality, the specific use of satellite data, but also what methods have been used to measure and analyse road traffic flow in the retrieved research. The information in the abstracts has also been used to achieve relevance ranking of the retrieved records to generate an elite set of documents for further reading.

These tools were provided to the client in different stages of completion during the project. While the question answering system lacks a graphical user interface and needs to be executed in a development environment, and consequently had to be run in our computer infrastructure, we made efforts to create a probabilistic search engine that the client could use to perform their own searches within the retrieved bibliographic material. Using openly available data provided by Crossref<sup>5</sup>, we also made it possible for the client to retrieve the relevant articles and download the full-text PDF file from the publisher.

Lastly, the search engine utilises the Altmetric web service<sup>6</sup> to display the Altmetric donut icon, including usage metrics of the research. By clicking on the icon, the client can retrieve usage data from various outlets, including mentions of the articles in patents, policy documents and newspapers. It also identifies mentions of the articles on social media platforms like Facebook and Twitter, as well as additional scientometric data available through the Dimensions citation database<sup>7</sup>, making it possible to explore citation data provided by Digital Science.

---

<sup>5</sup> <https://www.crossref.org/>

<sup>6</sup> <https://www.altmetric.com/>

<sup>7</sup> <https://www.dimensions.ai/>

## Question answering in abstracts

In order to enhance the possibilities to explore the datasets obtain from Web of Science, we have developed a strategy for extracting phrases from abstracts that potentially constitute answers to specific questions. This approach, known as question answering (QA), utilises deep learning and language modelling to identify sections in the texts that correspond to a maximum probability of constituting the answers to given questions. The language model used for this QA strategy was SciBERT (see Beltagy, Lo, & Cohan, 2019), which is trained on scientific papers available at the Semantic Scholar website. Examples of questions and answers extract from the air quality dataset are given below.

### **Q. What measurement methods and types of data sources are used to map and analyse air quality?**

**A.** CO<sub>2</sub> sensors, low-cost air quality sensors, optical SPM observations and meteorological measurements, GOCI and Himawari-8, IoT-based portable air quality measuring devices, aerosol laser ablation mass spectrometry, gas chromatography coupled with mass spectrometry, low-cost miniaturised FTIR spectrometers, open-path FTIR spectroscopy, low-dimensional Linear Ventilation Models, land-use regression (LUR) models, electronic nose

### **Q. How is satellite data used for air quality analyses?**

**A.** strengthen epidemiological studies investigating air pollution health effects, map and monitor surface air pollution, to reconstruct pollution concentrations at high spatio-temporal resolutions, spaceborne remote sensing, Land Surface Temperature (LST) and emissivity estimation, to investigate ground-level PM concentrations, estimate fine particle concentration on large spatial scale

With regard to the traffic safety dataset, the following question and examples of answers were produced:

### **Q. What methods are used to measure and analyse traffic flows in the road traffic network?**

**A.** average daily traffic, induction loop sensors, inertial sensors, loop detectors and radar sensors, lidar, longitudinal velocity and lateral acceleration variance, vibration-counting, magnetometer and accelerometer sensor, acceleration data from vehicle-mounted sensors, using features of the generated acoustic signals, acoustic measurements.

## Search engine for ranked retrieval of abstracts

In addition to the phrase extraction approach to answering natural language questions to material, a probabilistic search engine was developed within the project's scope. This search engine is an implementation of the BM25 best-match retrieval model (see Robertson & Zaragoza, 2009) and, as such, is based on the probability ranking principle of documents. This principle states that documents should be ranked in response to the user query according to the probability of the documents being relevant to the user (Robertson, 1977). The BM25 considers a binary property called *eliteness* of the relationship between document and terms, denoting the quality of a term being about the document content (Robertson & Zaragoza, 2009). This property is turn regarded to be statistically related to the local frequencies of the terms, as modelled by discrete Poisson distributions. To obtain a complete weighting scheme for the term-document relationship, the local frequencies are normalised by the length of the documents and multiplied by a weight conceptually similar to the inverse document frequency (IDF) weighting scheme introduced by Spärck Jones (1972).

In this project, the BM25 model was used to rank bibliographic records from Web of Science based on the content in the corresponding abstracts. Apart from ranking the records using the BM25 model, the search engine also offers the possibility to rank documents that are already selected by the BM25 document score according to citation frequency and citation frequency per year, respectively. To enhance the presentation of the records, an icon (called "donut") representing the Altmetric Attention Score is displayed for each document, containing information from Altmetric concerning the extent to which the document has been mentioned in sources like news articles, social media, Wikipedia, and patents. Screenshots of the search engine displaying results for the air quality and the traffic safety datasets, respectively, can be found in Figure 19 and Figure 20.

Figure 20 can serve as a use case for how the approaches presented in this report could be used together. The terms 'driver', 'drowsiness' were identified in the keyword map in Figure 16 and the co-word map in Figure 18. However, identifying specific terms is not enough. Finding the actual articles that discuss driver drowsiness makes it possible for the user to include this research in their work. The highlight feature helps the user evaluate if the results are relevant, while the possibility of switching to sorting the results based on citation data and the information provided by Altmetric further help evaluate if the article is relevant to use. Lastly, by clicking on the document's title, the user could find the article and download its PDF file.

**Using VIIRS Day/Night Band to Measure Electricity Supply Reliability: Preliminary Results from Maharashtra, India**

Michael L. Mann, Eli K. Melaas, Arun Malik

2016, REMOTE SENSING - Article

Unreliable electricity supplies are common in developing countries and impose large socio-economic costs, yet precise information on electricity reliability is typically unavailable. This paper presents preliminary results from a machine-learning approach for using satellite imagery of nighttime lights to develop estimates of electricity reliability for western India at a finer spatial scale. We use data from the Visible Infrared Imaging Radiometer Suite (VIIRS) onboard the Suomi National Polar Partnership (SNPP) satellite together with newly-available data from networked household voltage meters. Our results point to the possibilities of this approach as well as areas for refinement. With currently available training data, we find a limited ability to detect individual outages identified by household-level measurements of electricity voltage. This is likely due to the relatively small number of individual outages observed in our preliminary data. However, we find that the approach can estimate electricity reliability rates for individual locations fairly well, with the predicted versus actual regression yielding an  $R^2 > 0.5$ . We also find that, despite the after midnight overpass time of the SNPP satellite, the reliability estimates derived are representative of daytime reliability.

Score: 6.1808  
Citations: 13  
Citations/year: 2.60

**Spatiotemporal Prediction of Fine Particulate Matter During the 2008 Northern California Wildfires Using Machine Learning**

Colleen E. Reid, Michael Jerrett, Maya L. Petersen, Gabriele G. Pfister, Philip E. Morefield, Ira B. Tager, Sean M. Raffuse, John R. Balmes

2015, ENVIRONMENTAL SCIENCE &amp; TECHNOLOGY - Article

Estimating population exposure to particulate matter during wildfires can be difficult because of insufficient monitoring data to capture the spatiotemporal variability of smoke plumes. Chemical transport models (CTMs) and satellite retrievals provide spatiotemporal data that may be useful in predicting PM<sub>2.5</sub> during wildfires. We estimated PM<sub>2.5</sub> concentrations during the 2008 northern California wildfires using 10-fold cross-validation (CV) to select an optimal prediction model from a set of 11 statistical algorithms and 29 predictor variables. The variables included CTM output, three measures of satellite aerosol optical depth, distance to the nearest fires, meteorological data, and land use, traffic, spatial location, and temporal characteristics. The generalized boosting model (GBM) with 29 predictor variables had the lowest CV root mean squared error and a CV-R<sup>2</sup> of 0.803. The most important predictor variable was the Geostationary Operational Environmental Satellite Aerosol/Smoke Product (GASP) Aerosol Optical Depth (AOD), followed by the CTM output and distance to the nearest fire cluster. Parsimonious models with various combinations of fewer variables also predicted PM<sub>2.5</sub> well. Using machine learning algorithms to combine spatiotemporal data from satellites and CTMs can reliably predict PM<sub>2.5</sub> concentrations during a major wildfire event.

Score: 6.1282  
Citations: 96  
Citations/year: 16.00

**Enhancement of OMI aerosol optical depth data assimilation using artificial neural network**

A. Ali, S. E. Amin, H. H. Ramadan, M. F. Tolba

2013, NEURAL COMPUTING &amp; APPLICATIONS - Article

A regional chemical transport model assimilated with daily mean satellite and ground-based aerosol optical depth (AOD) observations is used to produce three-dimensional distributions of aerosols throughout Europe for the year 2005. In this paper, the AOD measurements of the Ozone Monitoring Instrument (OMI) are assimilated with Polyphemus model. In order to overcome missing satellite data, a methodology for preprocessing AOD based on neural network (NN) is proposed. The aerosol forecasts involve two-phase process assimilation and then a feedback correction process. During the assimilation phase, the total column AOD is estimated from the model aerosol fields. The main contribution is to adjust model state to improve the agreement between the simulated AOD and satellite retrievals of AOD. The results show that the assimilation of AOD observations significantly improves the forecast for total mass. The errors on aerosol chemical composition are reduced and are sometimes vanished by the assimilation procedure and NN preprocessing, which shows a big contribution to the assimilation process.

Score: 5.9829  
Citations: 6  
Citations/year: 0.75



Figure 19. Search result based on the air quality dataset.






driver drowsiness	
Highlight query terms <input checked="" type="checkbox"/> Sort by: Relevance	
<p><b>Detecting Driver Drowsiness Based on Sensors: A Review</b></p> <p>Arun Sahayadhas, Kenneth Sundaraj, Murugappan Murugappan</p> <p>2012, SENSORS - Review</p> <p>In recent years, <b>driver drowsiness</b> has been one of the major causes of road accidents and can lead to severe physical injuries, deaths and significant economic losses. Statistics indicate the need of a reliable <b>driver drowsiness</b> detection system which could alert the <b>driver</b> before a mishap happens. Researchers have attempted to determine <b>driver drowsiness</b> using the following measures: (1) vehicle-based measures; (2) behavioral measures and (3) physiological measures. A detailed review on these measures will provide insight on the present systems, issues associated with them and the enhancements that need to be done to make a robust system. In this paper, we review these three measures as to the sensors used and discuss the advantages and limitations of each. The various ways through which <b>drowsiness</b> has been experimentally manipulated is also discussed. We conclude that by designing a hybrid <b>drowsiness</b> detection system that combines non-intrusive physiological measures with other measures one would accurately determine the <b>drowsiness</b> level of a <b>driver</b>. A number of road accidents might then be avoided if an alert is sent to a <b>driver</b> that is deemed drowsy.</p>	<p>Score: 11.3353 Citations: 240 Citations/year: 26.67</p> 
<p><b>Driver drowsiness recognition via transferred deep 3D convolutional network and state probability vector</b></p> <p>Lei Zhao, Zengcai Wang, Guoxin Zhang, Huanbing Gao</p> <p>2020, MULTIMEDIA TOOLS AND APPLICATIONS - Article</p> <p><b>Driver drowsiness</b> is a major cause of road accidents. In this study, a novel approach that detects human <b>drowsiness</b> is proposed and investigated. First, <b>driver</b> face and facial landmarks are detected to extract facial region from each frame in a video. Then, a residual-based deep 3D convolution neural network (CNN) that learned from an irrelevant dataset is constructed to classify <b>driver</b> facial image sequences with a certain number of frames for obtaining its <b>drowsiness</b> output probability value. After that, a certain number of output probability values is concatenated to obtain the state probability vector of a video. Finally, a recurrent neural network is adopted to classify constructed probability vector and obtain the recognition result of <b>driver drowsiness</b>. The proposed method is tested and investigated using a public drowsy <b>driver</b> dataset. Experimental results demonstrate that similar to 2D CNN, 3D CNN can learn spatiotemporal features from irrelevant dataset to improve its performance obviously in <b>driver drowsiness</b> classification. Furthermore, the proposed method performs stably and robustly, and it can achieve an average accuracy of 88.6%.</p>	<p>Score: 11.1282 Citations: 0 Citations/year: 0.00</p> 
<p><b>A Real-time Driving Drowsiness Detection Algorithm With Individual Differences Consideration</b></p> <p>Feng You, Xiaolong Li, Yunbo Gong, Haiwei Wang, Hongyi Li</p> <p>2019, IEEE ACCESS - Article</p> <p>The research work about driving <b>drowsiness</b> detection algorithm has great significance to improve traffic safety. Presently, there are many fruits and literature about driving <b>drowsiness</b> detection method. However, most of them are devoted to find a universal <b>drowsiness</b> detection method, while ignore the individual <b>driver</b> differences. This paper proposes a real-time driving <b>drowsiness</b> detection algorithm that considers the individual differences of <b>driver</b>. A deep cascaded convolutional neural network was constructed to detect the face region, which avoids the problem of poor accuracy caused by artificial feature extraction. Based on the Dlib toolkit, the landmarks of frontal <b>driver</b> facial in a frame are found. According to the eyes landmarks, a new parameter, called Eyes Aspect Ratio, is introduced to evaluate the <b>drowsiness</b> of <b>driver</b> in the current frame. Taking into account differences in size of <b>driver's</b> eyes, the proposed algorithm consists of two modules: offline training and online monitoring. In the first module, a unique fatigue state classifier, based on Support Vector Machines, was trained which taking the Eyes Aspect Ratio as input. Then, in the second module, the trained classifier is application to monitor the state of <b>driver</b> online. Because the fatigue driving state is gradually produced, a variable which calculated by number of drowsy frames per unit time is introduced to assess the <b>drowsiness</b> of <b>driver</b>. Through comparative experiments, we demonstrate this algorithm outperforms current driving <b>drowsiness</b> detection approaches in both accuracy and speed. In simulated driving applications, the proposed algorithm detects the drowsy state of <b>driver</b> quickly from 640x480 resolution images at over 20fps and 94.93% accuracy. The research result can equip intelligent transportation system, ensure <b>driver</b> safety and</p>	<p>Score: 11.1203 Citations: 1 Citations/year: 0.50</p> 

Figure 20. Search result based on the traffic safety dataset. Here, the highlight feature is turned on to show the search terms used.

## Conclusions

This report has served to describe the means of providing tools and opportunities to investigate and evaluate research on the subject matters in focus for the project. It has also allowed us to develop approaches for creating a framework for doing scientometric and machine learning-based text analysis on research articles and their content. Furthermore, it has served as a stepping stone for developing combined scientometric and semantic machine learning approaches. In research terminology, this entails methods development that can yield new insights into bibliographic data that will be generalised in further publications from the Data as Impact Lab.

A second purpose of the development of these tools was to showcase the use of machine learning and, in a sense, AI technology to analyse text in large bibliographic data sets. Time did not permit us to develop any language comprehension models of our own. Nevertheless, using pre-trained language models and developing a probabilistic search engine made it possible to identify specific techniques and methodologies used within the research.

While exploratory, we show that in conjunction with the insights gained from the scientometric approach, which provides the opportunity for the user to identify topics and possibly to formulise hypotheses about the content of the retrieved research, using terms and phrases identified in the visualisations, the user can ask questions to the material in the Question Answer system. Using insights from the visualisations and the answers to the questions posed, the user can use the search engine to identify the particular documents wherein the methodologies and techniques have been applied.

Taken together, these approaches provide a complete ecosystem for the user who wants to explore the research about a subject matter and to identify relevant research covering specific aspects of interest within the large set of results retrieved through the Boolean searches described at the beginning of the report.

## References

- Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A Pretrained Language Model for Scientific Text. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. <https://doi.org/10.18653/v1/d19-1371>
- Moretti, F. (2013). *Distant Reading*. London: Verso.
- Persson, O. (1994). The intellectual base and research fronts of JASIS 1986–1990. *Journal of the American Society for Information Science*, 45(1), 31-38.
- Robertson, S. (1977). The probability ranking principle in IR. *Journal of Documentation*, 33(4), 294–304. <https://doi.org/10.1108/eb026647>.
- Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4), 333–389. <https://doi.org/10.1561/15000000019>
- Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1), 11–21. <https://doi.org/10.1108/eb026526>
- Van Eck, N.J., & Waltman, L. (2010). Software survey: *VOSviewer*, a computer program for bibliometric mapping. *Scientometrics*, 84(2), 523-538.